Visual and Auditory Factors Facilitating Multimodal Speech Perception

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation with distinction in Speech and Hearing Sciences in the undergraduate colleges of The Ohio State University

by

Pamela J. Ver Hulst

The Ohio State University
June 2006

Project Advisor: Dr. Janet Weisenberger, Department of Speech and Hearing Science

Abstract

Speech perception is often described as a unimodal process, when in reality it involves the integration of multiple sensory modalities, specifically, vision and hearing. Individuals use visual information to fill in missing pieces of auditory information when hearing has been compromised, such as with a hearing loss. However, individuals use visual cues even when auditory cues are perfect, and cannot ignore the integration that occurs between auditory and visual inputs when listening to speech.

It is well known that individuals differ in their ability to integrate auditory and visual speech information, and likewise that some individuals produce clearer speech signals than others, either auditorily or visually. Clark (2005) found that some talkers in a study of the McGurk effect, produced much stronger 'integration effects' than did other talkers. One possible underlying mechanism of auditory + visual integration is the substantial redundancy found in the auditory speech signal. But how much redundancy is necessary for effective integration? And what auditory and visual characteristics make a good integration talker?

The present study examined these questions by comparing the auditory intelligibility, visual intelligibility, and the degree of integration for speech sounds that were highly reduced in auditory redundancy, produced by 7 different talkers. Performance of participants under four conditions: 1) degraded auditory only, 2) visual only, 3) degraded auditory + visual, and 4) non-degraded auditory + visual, was examined. Results indicate across-talker differences in auditory and auditory + visual intelligibility. Degrading the auditory stimulus did not affect the overall amount of McGurk-type integration, but did influence the type of McGurk integration observed.

Acknowledgments

I would like to thank and acknowledge Dr. Janet Weisenberger for giving me the opportunity to work with her on this research study, as well as the support and guidance that she has given me throughout the entire process. I would like to thank Natalie Feleppelle for all the time, advice, and assistance that she has contributed. I would also like to thank my co-laborer, Elizabeth Anderson for the help she provided when I needed her most. I would like to extend my gratitude and appreciation to the participants of my study, who put forth the time that made the outcome of this study possible. In addition I would like to thank my family and my boyfriend, who gave me constant support and encouragement.

The present study was supported by an ASC Undergraduate Research Scholarship and by the SBS Undergraduate Research Scholarship.

Table of Contents

Chapter 1: Introduction and Literature Review

Speech is the integrated use of multiple sensory modalities; auditory and visual. A listener may perceive speech as being purely auditory, but when the talker can be seen, visual information is incorporated with auditory cues so the listener can fully understand the intent of the information being conveyed. Visual information has been proven to be useful in environments where auditory cues are not sufficient (noisy environments, loss of hearing, etc.). When a talker's face is visible in a noisy environment, the intelligibility of the auditory speech is notably better than auditory alone speech perception (Munhall, 2002). However, McGurk and MacDonald (1976) showed that visual information plays a role even in the perception of clear, unambiguous speech tokens. They demonstrated this by dubbing a set of auditory syllables such as the bilabial consonant (ba) onto video recording of a speaker saying velar consonants, such as (ga), which when integrated, produces a perception of the fusion of the two, or (da). When the listener/observer looks away from the video screen, the auditory tokens are heard correctly. The integration of the two modalities in this case has been termed "fusion tokens." The consonant-vowel (ba) is made by air being pushed up through the glottis stopping at the lip articulators and then making a burst, which then continues into the open (ah) sound. The consonant-vowel formation of (ga) is produced in a similar manner, but at a different articulator and position of the mouth; it is made at the back of the mouth on the velum, which is difficult for an observer to see. As a result, when the two places of articulation are integrated together (da) is perceived, which is made between the lips and the

velum at the alveolar ridge, in between the lips and the velum.

Combination of two stimuli was also observed by McGurk and MacDonald. When (ga) was heard, while (ba) was seen on the videotape, the observer perceived the sound to be (bga), a combination. The bilabial formation of the (ba) was clearly seen on the screen, while the observer was hearing a (ga), which led to the perception of a combination, (bga). In both instances, where there was either fusion or combination, the two stimuli were integrated together to form a new phoneme. This provides evidence that speech perception is multimodal and not just auditory.

There are several theories that have attempted to explain the process of audio-visual speech perception. The single channel theory states that only one modality is necessary to perceive speech and to identify a speech sound. While this may be true in some circumstances, the McGurk effect provides evidence that two modalities are present in the perceptual process even if the auditory signal is perfect. The multichannel theory states that the audio and visual are processed separately and are only integrated when the listener is looking for a response. If the two modalities are different, then the listener is able to give a response that accommodates the difference. This theory does accommodate phenomena such as the McGurk effect. This brings up the question of when the two modalities are integrated in the perception process.

Early integration is the idea that information from the visual and auditory is integrated before a decision is concluded, resulting in a single decision (Robert-Ribes, Schwartz, and Escudier, 1995). Green and Kuhl (1988) suggested that

integration happens before phonemic categorization, which means that visual and auditory information is combined prior to the time decisions are made. Their study investigated whether other dimensions of the auditory signal, such as voice-onset-time (VOT), affected the resulting McGurk effect. What they found was that the perception of voicing could be influenced by the visual modality. Because visual information alone is inadequate to allow a decision on VOT, they concluded that the decision must be dependent upon both modalities, which are combined by the time a phonetic decision is made. This rules out the possibility that there is a post-phonetic integration, or late integration.

The Fuzzy Logic Model of Perception (Massaro, 1998) suggests that the visual and auditory information are mapped onto the prototype at the same level of phonetic processing, which can also be called late-integration. Although there is early interaction among the visual and auditory modalities, integration of the information does not occur until late in the perception process. Such a process could also accommodate phenomena such as the McGurk effect.

McGurk effects have been obtained in different situations, whether with adults, children, or speakers of other languages. The McGurk effect has been documented to appear in infants as young as five months old (Burnham & Dodd, 1996; Desjardins & Werker, 1996; Roenblum, Suchmuckler & Johnson, in press). As age progresses it has been found that the strength of the McGurk effect increases. McGurk and MacDonald (1976) found that the effect was greater in 7-8 year olds than 3-5 year olds, and even greater in adults, which supports the idea of experiential learning. There have been limited studies on the McGurk

effect cross culturally. Studies have found that there is a weaker McGurk effect in Japanese participants than American (Sekiyama and Tohura, 1993). This incongruence is attributed to the difference in cultures. Compared to the American culture, Japanese speakers tend to make less eye contact when engaged in conversation. The Japanese speakers are receiving equal exposure to auditory signals, but unequal exposure to visual signals. This results in the difference between the two cultures and provides information that the perception of speech is multimodal.

Talker characteristics are made up of auditory and visual cues that provide information during speech perception that has little variation across cultures. Auditory cues provide information about the place, manner, and voicing of a phoneme when produced. The place of articulation is where the articulation is being produced. Articulation can be made by forming bilabials (on the lips), labiodentals (lower lip and upper front teeth), interdentals (the tongue and teeth), alveolar (tongue tip and alveolar ridge), palatal-alveolar (tongue blade and alveolar ridge), palatals (tongue and hard palate), and velars (tongue and soft palate). The manner of articulation is the behavior the phoneme carries out when it is being produced. This is whether it is a stop, a fricative, affricate, liquid, or glide. Lastly, consonants can either be voiced or voiceless. Sounds that are produced when the vocal folds are vibrating are voiced, such as in the phoneme /p/. Nasal stops and vowels are always voiced phonemes. When the vocal folds are not vibrating, they are said to be voiceless sounds, such as in the phoneme /b/.

When describing vowels, there are three main components: 1) vowel height, 2) backness, and 3) the degree of lip rounding. A description of two vowels are in the example following; /i/ is a vowel is a high vowel, made in the front of the mouth, with lip spreading (little lip rounding), while the vowel /o/ is a high-middle vowel made toward the back of the mouth, with a good degree of lip rounding.

Speech is also characterized by visual cues. The phoneme is a unit of speech that can be used for writing a language down in a systematic and unambiguous manner. The phoneme is a family of a variation of sounds, but is still recognized as the same linguistic unit. In the same way, some speech sounds are similar enough in their visual characteristics to be considered single units (Jackson, 1988). The sounds, although different, are described as possessing the same visual characteristics. These sounds that have been grouped together are defined as being a visual phoneme, or more commonly known as a viseme. Visemes usually contains more than one speech sound and within each, the speech sounds are produced with similar movement patterns. Visemes exist for both consonants and vowels. /p, b, m/ are all bilabial consonants, that are different auditorily, but the same when compared visually. This is why it is difficult for those who are learning to lip-read. Auditory cues are conveyed by the place, manner, and voicing of a phoneme, while visually the observer only knows the place of articulation, but even that can sometimes be very ambiguous. The /k/ and /g/ are difficult to see since they are made in the back of the mouth, but the difference is not hard to hear. Together auditory and

visual cues are integrated to make clear, unambiguous speech, which renders conversational speech more intelligible in compromised environments.

Studies have shown that “clear speech” is significantly more intelligible than conversational speech for hearing-impaired listeners within a quiet background as well as for normal and hearing-impaired listeners within a noisy background (Picheny *et al.,* 1985; Uchanski *et al.,* 1996; Payton *et al.,* 1994). Clear speech contains spectral and temporal characteristics that make it highly intelligible.

There are numerous benefits of auditory-visual speech perception over listening alone or speechreading alone. The addition of visual cues causes an increase in the speech-to-noise ratio by 15 dB (Sumby and Pollack, 1954). Depending by the speech of the talker, each 1 dB improvement in S/N can correspond to a 5 to 10 percentage point increase in intelligibility (Miller *et al.,* 1951; Grant and Braida, 1991). The addition of speechreading can determine the difference between perfect comprehension and failure to understand.

The improvement of speech understanding in noisy environments can be explained with three possible roles by Summerfield (1987). First, speechreading provides segmental (vowel and consonant), and suprasegmental (intonation, stress, etc.) information. This is redundant in that acoustic cues also provide segmental and suprasegmental information. Segmental information is conveyed by the measurement of the mouth opening and the sound that is produced. For example, vowels are defined by the height and spread of the mouth, while consonants use a variety of articulators that vowels do not, such as the teeth and

alveolar ridge. This information can also be heard acoustically, and be distinguished determined on the sound that is produced. Suprasegmental information can be conveyed through how sounds are spoken, such as how the word is stressed. This can be heard, but seen as well by the emphasis that is put on certain sounds, which can be seen at articulating points. Second, speechreading provides segmental and suprasemental information complementary to cues provided acoustically. There are some acoustic cues that are not available to the listener in some instances, such as a noisy environment that are conveyed by the visual cues. Acoustic cues for place of articulation helps the listeners distinguish between consonants, but are also somewhat easy to speechread. This is fortunate, since these acoustic cues are the first to be lost to noise. Last, when a listener watches a talker speak, the acoustic speech signal and the visible movement of the talker's lips share common spatial and temporal properties. The listener can use this commonality to direct themselves to the speech signal of interest rather than to any surrounding background noise. There are some auditory cues that can not be determined with visual cues, such as the voicing of some consonants.

Massaro (1998) states that "auditory-visual integration" is the process employed by individual receivers to combine information extracted from auditory and visual sources. Integration of auditory and visual cues is thus distinct from the ability to extract auditory and visual cues and high-order language processing of the information received by the two senses. Grant and Seitz (1998), questioned whether individual listeners integrate auditory and visual cues with

varying degrees of efficiency. From the study it was determined that auditory-visual speech integration is a measurable skill the subjects use whenever auditory and visual sources of information are available (these results are independent from auditory and visual cue peripheral cue extraction). It is assumed that auditory-visual speech integration implies better speech recognition performance than from auditory-alone or visual-alone. This assumption is with the exception of artificially created stimulus situations, such as studies of the McGurk effect. Grant and Seitz found that older subjects tended to be poorer integrators than younger subjects. Braida (1991) suggests that integration efficency may be treated as separate from the extraction of sensory information, which results in making it possible to have efficient integration and still perform very poorly on speech recognition tasks. This may happen when auditory and/or visual information is lacking, in situations, for example, where the subjects may be deaf or hard-of-hearing. Theoretically, it is possible to extract the sufficient auditory and visual cues but still do poorly on auditory-visual speech perception tests because of underprivileged integration skills.

In a previous study in our laboratory (Clark, 2005), there was variability in the degree to which subjects exhibit the McGurk effect, and it was questioned whether subjects show a reduced McGurk response when viewing self as talker. While the data showed that half of the subjects did show a reduced McGurk effect to themselves as a talker, further analysis indicated that none of the subjects showed a strong McGurk effect to these particular talkers. This result suggests that there may be particular talker characteristics that influence the

occurrence of audiovisual speech perception. Braida found that talkers tend to have different strategies of how to produce clear and intelligible speech at different speaking rates. To determine what characteristics make a “good” talker in any speaking rate, it needs to be known whether the observer needs to have both auditory and visual cues to perceive the talker as having clear, intelligible speech, or whether there can be a missing cue. If both cues are required, it needs to be determined whether there can be some ambiguity in one, or both of the cues. Lastly, it needs to be determined what physical characteristics of both the visual and auditory cues make the speaker a more intelligible talker.

The present study investigated talker characteristics that produce good auditory perception, good visual perception, and good audio-visual perception. Specifically, the research examined whether persons who are good auditory talkers, or persons who are easy to speech read, also produce high levels of audio-visual integration. A series of talkers were video recorded producing single syllable speech tokens. These recordings were digitally edited and presented to a group of subjects under audio only, visual only and audio-visual conditions. To avoid ceiling-effects, the speech samples were degraded by phase inverting 50 percent of the samples in the speech waveform, effectively reducing the redundancy of the auditory samples without adding extraneous noise. Integration was determined by 1) comparing listeners’ abilities in degraded and non-degraded auditory + visual conditions to degraded auditory only and visual only conditions, and also 2) measuring listeners’ responses to degraded auditory + visual McGurk stimuli.

Chapter 2: Method

Participants

Participants included seven talkers and ten observers. The talkers consisted of 3 female and 4 male participants with ages ranging from 20 to 23, who produced a set of eight single syllable stimuli that were recorded by a video camera. All the talkers were undergraduate university students and had reported having normal hearing and normal or corrected vision. The observers consisted of 8 female and 2 male participants with ages ranging from 19 to 26. Nine of the ten were undergraduate university students in the Speech and Hearing Sciences major, and one was a graduate student with focus on Speech-Language Pathology. All ten observers also reported having normal hearing and normal or corrected vision. None of the participants reported knowing about the McGurk effect.

Interfaces for Stimulus Presentation

- Visual Apparatus: 20 inch video monitor
- Auditory Apparatus: Sennheiser circumaural headphones

Stimuli

A limited set of eight nonsense syllables was used in this study. These syllables were chosen specifically to satisfy the following conditions:

1. Pairs of stimuli were minimal pairs, differing by only the initial phoneme

2. All stimuli were accompanied by the vowel /ae/ because it does not involve lip rounding or lip extension.

3. There were multiple stimuli in each category of articulation: place (bilabial, alveolar), manner (stop, fricative, nasal), and voicing (voiced, voiceless).

4. All stimulus were presented without a carrier phrase

The eight single-syllable stimuli (undubbed) used were as follows:

Bilabial	mat, bat, pat
Alveolar	sat, zat, tat
Velar	gat, cat

The four following (dubbed) dual-syllable stimuli were used (the beginning column shows the auditory stimulus, which was dubbed onto the visual stimulus).

1. bat-gat
2. gat-bat
3. cat-tat
4. tat-cat

Presentation Conditions

Visual only single-syllable
Auditory only (Degraded) single-syllable
Audio-Visual (Degraded) single-syllable
Audio-Visual (Non-Degraded) single-syllable
Audio-Visual (Degraded) dual-syllable
Audio-Visual (Non-Degraded) dual-syllable
60 stimuli per condition per talker

Stimulus Recording and Editing

Syllables from seven different talkers were recorded onto digital videotape to provide the visual stimuli. Auditory recordings were made from

direct microphone input to the computer with a 48 kHz sampling rate. Auditory stimuli were degraded using a speech software program that randomly phase-inverted 50 percent of the samples from the speech waveform. This process disrupts the spectral fine structure of the waveform, but preserves the temporal envelope. In this way, redundancy is reduced without adding noise to the waveform. Audio-visual stimuli were made using a digital video editing program that allowed degraded auditory stimuli to be dubbed onto visual stimuli.

Procedure

The testing was conducted individually in a sound-booth with the lights turned off so that viewer had optimal viewing of the television set through the sound booth window, where light from the exterior room came through. The participant sat comfortably and was presented the visual stimuli on a DVD player that was connected to a 20 in. video monitor, placed on a stand at eye level. The monitor was approximately 1.5 meters from the participant with the auditory stimulus being presented through Sennheiser circumaural headphones at a comfortable volume. Each participant watched 28 videos, 4 videos per talker that was video recorded, each containing sixty trials in different visual and auditory conditions. Participants were instructed to verbally respond to what they perceived on the video monitor, and/or the headphones, while an experimenter transcribed their responses. Participants were informed that they would encounter both words, and non-sense words, including phoneme sequences they may not encounter in the English language.

Chapter 3: Results and Discussion

Single-Syllable Stimuli

The first step in the analysis of the data was to look at the single-syllable stimuli to determine the percent correct responses in identification for degraded auditory only, visual only, and audio-visual presentation conditions. Responses were transcribed as correct when the observer responded with the syllable that had actually been produced by the talker. Figures 1, 2, and 3 show performance averaged across observers for each of the seven talkers in the auditory only, visual only, and audio-visual conditions, respectively.

Observers showed better performance in the auditory only condition than in the visual only condition. In the audio-visual condition observers reached higher percent correct performance than in either auditory only or visual only. This suggests that observers are able to integrate auditory and visual cues to improve the intelligibility of the stimulus.

Examination of performance produced by specific talkers in Figures 1-3 shows that talkers 2, 5, and 6 yielded the highest levels of auditory intelligibility, as well as the highest levels of audio-visual integration. Interestingly, talkers 2 and 5 show the lowest levels of visual intelligibility. Thus, it appears that good auditory intelligibility, combined with a certain degree of visual ambiguity, may produce better integration. However, given that the differences in visual intelligibility across talkers are not great, it is equally likely that performance in the audio-visual condition is determined primarily by auditory intelligibility.

Statistical analysis using a one-factor (talker), repeated measures analysis

of variance indicated that the differences across talkers in auditory intelligibility shown in Figure 1 were significant [$F(6,63) = 14.04$, $p<.05$]. Similarly, significant differences across talkers were found in the audio-visual condition [$F(6, 63) = 7.25$, $p<.05$], but not in the visual only condition [$F(6, 63) = 1.57$, ns].

To determine whether talker intelligibility in the auditory only condition was related to talker intelligibility in the visual only condition, the Pearson r correlation coefficient was calculated. No significant relationship was found [$r = -.16$], indicating that the talkers with the best auditory intelligibility were not necessarily those with the best visual intelligibility. However, a significant, moderately sized correlation was observed for auditory only and audio-visual performance across talkers [$r = .44$, $p<.05$]. More surprisingly, a significant negative correlation between visual only and audio-visual performance was found [$r = -.62$, p,.05]. This negative relationship could be interpreted to suggest that poor visual intelligibility was more likely to lead to integration of auditory and visual inputs.

Dual-Syllable Stimuli

The second step in the analysis of the data was to look at the dual-syllable stimuli, in which one syllable was presented via the auditory channel and a different syllable was presented via the visual channel. These stimuli provide an opportunity for observers to exhibit McGurk-type integration of the inputs. Note that there is no “correct” response for these stimuli. Instead, responses are categorized as “auditory” when the participant responded with the auditory stimulus, “visual” when the participant responded with the visual stimulus, and

"other" when the participant gave a response that was neither the auditory nor the visual stimulus. These "other" responses can be thought of as indicating integration of the visual and auditory inputs.

Figure 4 shows response categories for each of the seven talkers for the normal and degraded audio-visual presentation conditions. As can be seen, observers showed a heavy reliance on the auditory input when it was not degraded, with "auditory" responses considerably more frequent than when the input was degraded. Conversely, Figure 5 indicates that these same observers relied much more heavily on the visual input when the auditory input was degraded, with a much greater incidence of "visual" responses. Finally, the proportion of "other" responses is shown in Figure 6. Although the pattern of visual or auditory reliance, as shown in Figures 4 and 5, was very different for the two presentation conditions, no single pattern for the percentage of "other" responses was found across talkers. For talkers 1, 2, 3, and 7, it appears that there is a slightly higher percentage of "other" responses for the degraded auditory condition; for talker 5 the opposite result is true; and for talkers 4 and 6 the two presentation conditions produced approximately equal percentages of "other" responses.

Statistical analysis using a dependent groups t-test supported these observations. Significantly more "auditory" responses were observed for the normal auditory condition, as compared to the degraded auditory condition [$t(6) = 8.21$, $p<.05$], and significantly more "visual" responses were observed for the degraded auditory condition [$t(6) = -10.03$, $p<.05$]. However, no significant

differences were found for the percentage of “other” responses.

Types of Integration Responses

Finally, the percentage of “other” responses was examined in detail, to determine whether degraded and non-degraded auditory presentation produced differences in the type of integration that participants exhibited. Figures 7 and 8 show a breakdown of these responses across talkers. Responses are classified as a “fusion” (e.g., when the observer “averages” the place of articulation of the initial consonant, such as when an auditory /bat/ and a visual /gat/ produce a response of /dat/.); a “combination” (e.g., when an auditory /gat/ and a visual /bat/ produce a response of /bgat/); or “neither” (a different response entirely).

In Figure 7, which shows performance for the degraded auditory condition, it can be seen that a preponderance of “fusion” responses is shown for all talkers. Only a tiny percentage of “combination” responses is seen, and low to moderate levels of “neither” responses. Figure 8 shows the same categorization for the normal auditory presentation condition. Here the pattern is a bit more complex. For talkers 1, 2, 3, and 7, there is again a substantial percentage of “fusion” responses. However, for talkers 4, 5, and 6, this is not evident. For these talkers, the percentage of traditional McGurk-type integration is actually greater for the degraded auditory input than for the normal auditory input. Dependent groups t-tests comparing the percentage of fusion, combination, and neither responses across presentation conditions did not indicate significant differences in performance. However, these analyses averaged all talkers

together, which might have obscured possible cross-talker differences.

Chapter 4: Summary and Conclusion

Overall, it appears that there are differences across talkers in how well they produce audio-visual stimuli that can be integrated by observers. In the percent correct identification performance analysis, talkers 2, 5, and 6 produced the highest levels of audiovisual integration. They also produced the highest levels of auditory only identification. This result suggests that greater auditory intelligibility promotes audiovisual integration. The auditory intelligibility was rather still intelligible, suggesting that the stimuli were not effectively reduced in redundancy. Further studies may need to examine whether even less intelligibility in the auditory signal produces the same results that were produced in the present study.

However, for the dual-syllable stimuli, two of these same talkers (talkers 5 and 6) produced a greater percentage of typical McGurk-type fusion integration in the degraded auditory condition, as compared to the normal auditory condition. This result, conversely, suggests that a certain degree of auditory ambiguity promotes audiovisual integration.

The present results represent only a beginning look at the question of whether there are “good” and “poor” talkers for audiovisual integration. Follow up work is needed to examine more closely the characteristics of good and poor talkers. First, auditory spectral analyses of specific talker utterances should be analyzed to determine whether tokens produced by good talkers are closer to prototypical consonant or vowel productions, or are more similar to clear speech tokens. Second, the visual characteristics of talkers need to be evaluated.

Vatikiotis-Bateson and colleagues have provided some metrics for such evaluation in providing optimal parameters for speechreading.

Again, the question to be addressed is whether a certain amount of ambiguity in either the auditory or the visual input is more conducive to audiovisual integration, or whether clarity in at least one input channel produces better integration.

The results of the present study have long-term implications for the development of aural rehabilitation programs for individuals with hearing impairments. If the parameters that lead to optimal audiovisual integration can be identified, then training materials utilizing these optimized parameters can be produced to provide initial training for patients.

Chapter 5: References

Burnham, D. & Dodd, B. (1996). Auditory-visual speech perception as an automatic process: The fusion effect in human infants and across languages. In D. Stork & M.E. Hennecke (Eds), *Speech Reading by Humans and Machines.* Berlin: Springer-Verlag.

Clark, C. (2005). *Effects of Long Term Audio-Visual Versus Audio-only Experience on Multimodal Speech Integration.* Senior Honors Thesis. The Ohio State University.

Grant, K.W. & Braida, L.D. (1991). Evaluating the Articulation Index for audiovisual input, *J. Acoustical Society of America*. 89, 2952-2960.

Grant, K.W., & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences, *Journal of the Acoustical Society of America, 104,* 2438-2450.

Jackson, P. (1988). *The Theoretical Minimal Unit for Visual Speech Perception: Visemes and Coarticulation.* The Volta Review. 98-114.

Massaro, D.W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle.* Cambridge, MA: MIT Press.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature,* 264, 746-748.

Munhall, K.G., and Vatikiotis-Bateson. *Spatial and Temporal Constraints on Audiovisual Speech Perception.*

Payton, K.L., Uchanski, R.M., and Braida, L.D. (1994). *Intelligibility of conversational and clear speech in noise and reverberation for listeners*

with normal and impaired hearing. J. Acoust. Soc. Am. 95, 1581-1592.

Picheny, M.A., Durlach, N.I., and Braida, L.D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech, *Journal Speech Hear. Res. 28*, 96-103.

Robert-Ribes, J., Schwartz, J.L. & Escudier, P. (1995). "A comparison of modals for fusion of the auditory and visual sensors in speech perception." *Artificial Intelligence Review, 9,* 323-46.

Rosenblum, L.D., Schmuckler, M.A. & Johnson, J.A. (1997). The McGurk effect in infants. *Perception and Psychophysics, 59,* 347-57.

Sekiyama, K. & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics, 21,* 427-44.

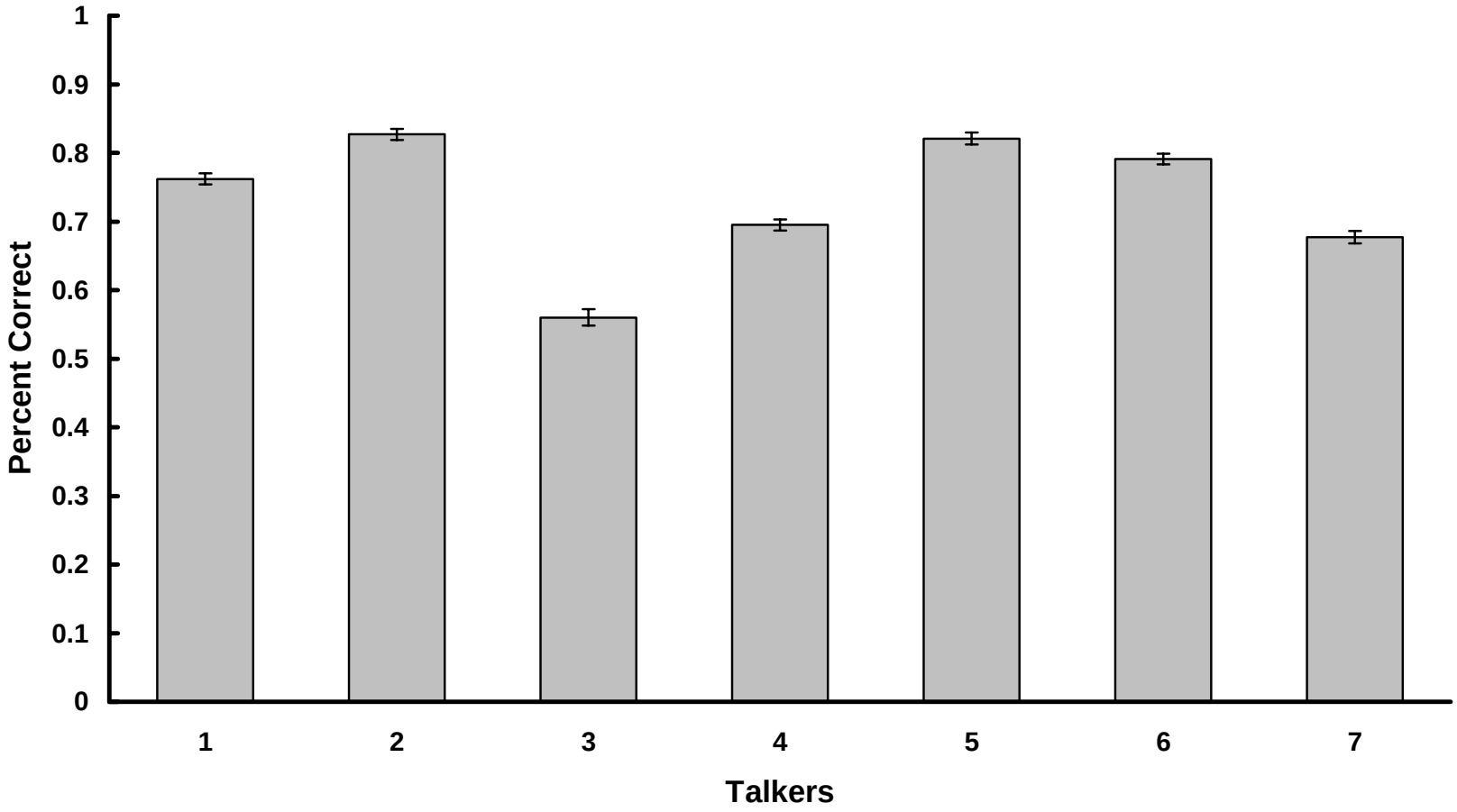
Sumby, W.H. & Pollack, I. (1954). Visual contributions to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212-15.

Summerfield, A.Q. (1987). Some preliminaries to comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds), *Hearing by Eye: The Psychology of Lipreading.* Hove, UK: Lawrence Erlbaum Associates Ltd.

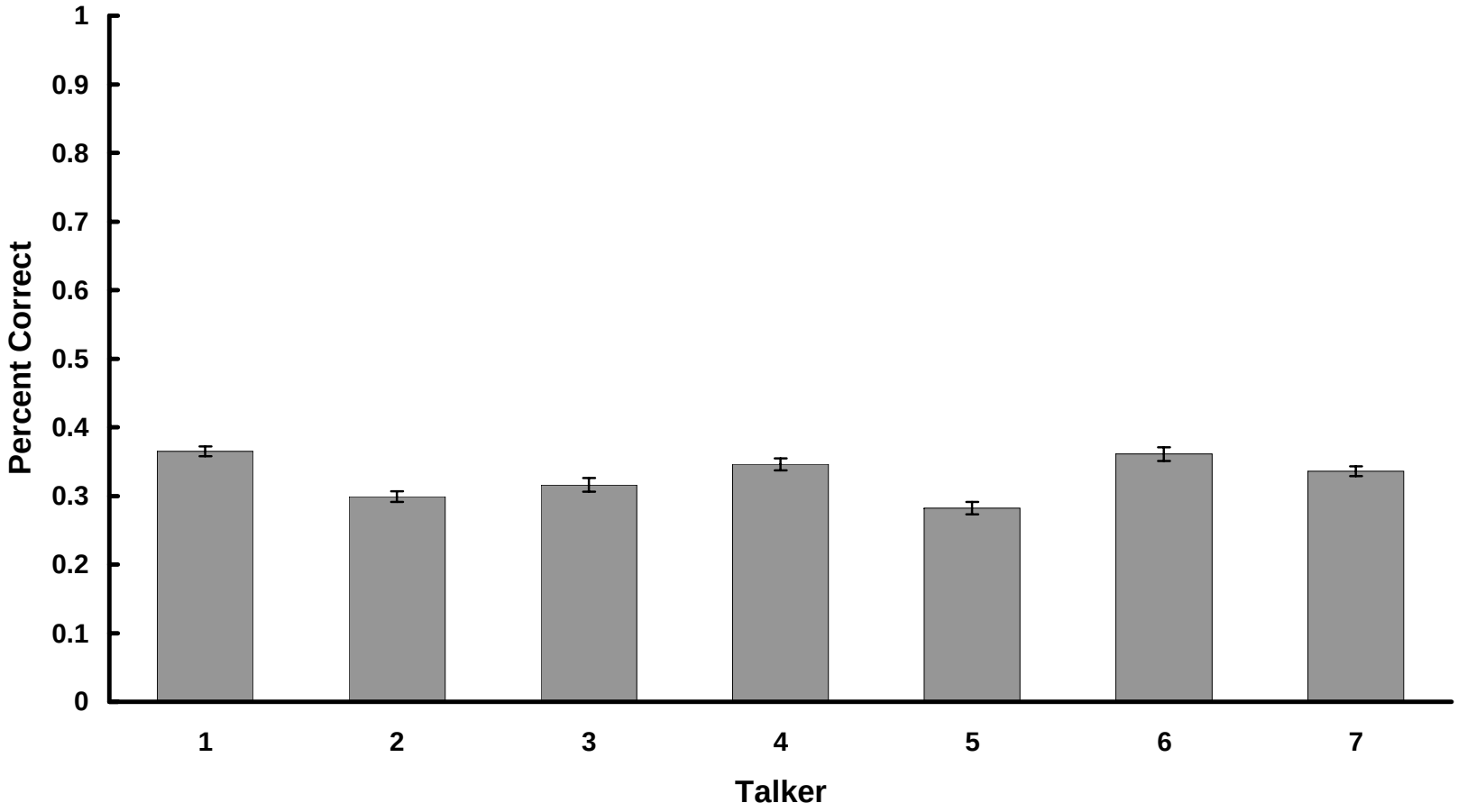
Uchanski, R.M., Choi, S., Braida, L.D., Reed, C.M., and Durlach, N.I. (1996). *Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate*, J. Speech Hearing Research. 39, 494-509.

List of Figures

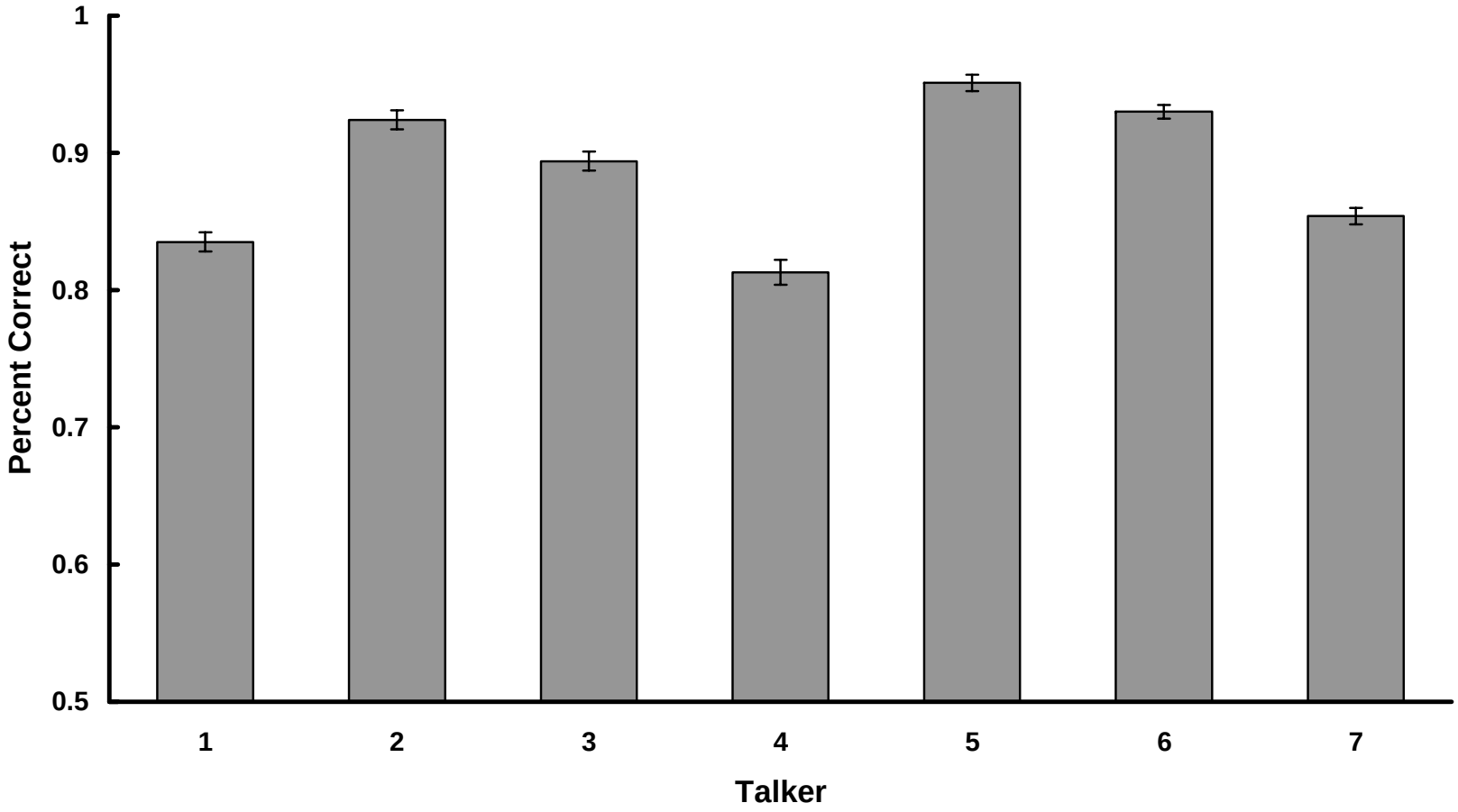
Percent Correct Identification Auditory --Degraded
Figure 1.



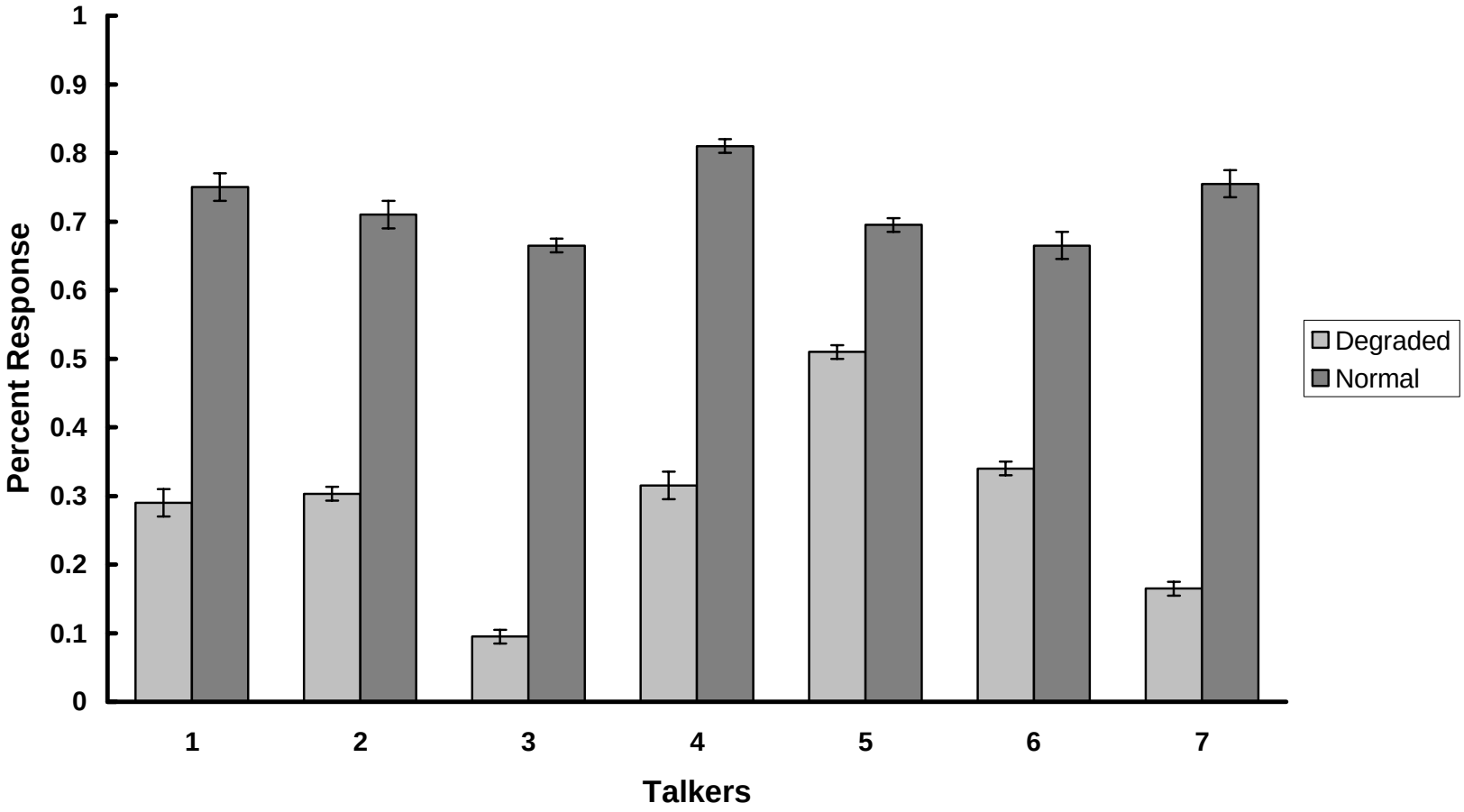
Percent Correct Identification Visual
Figure 2.



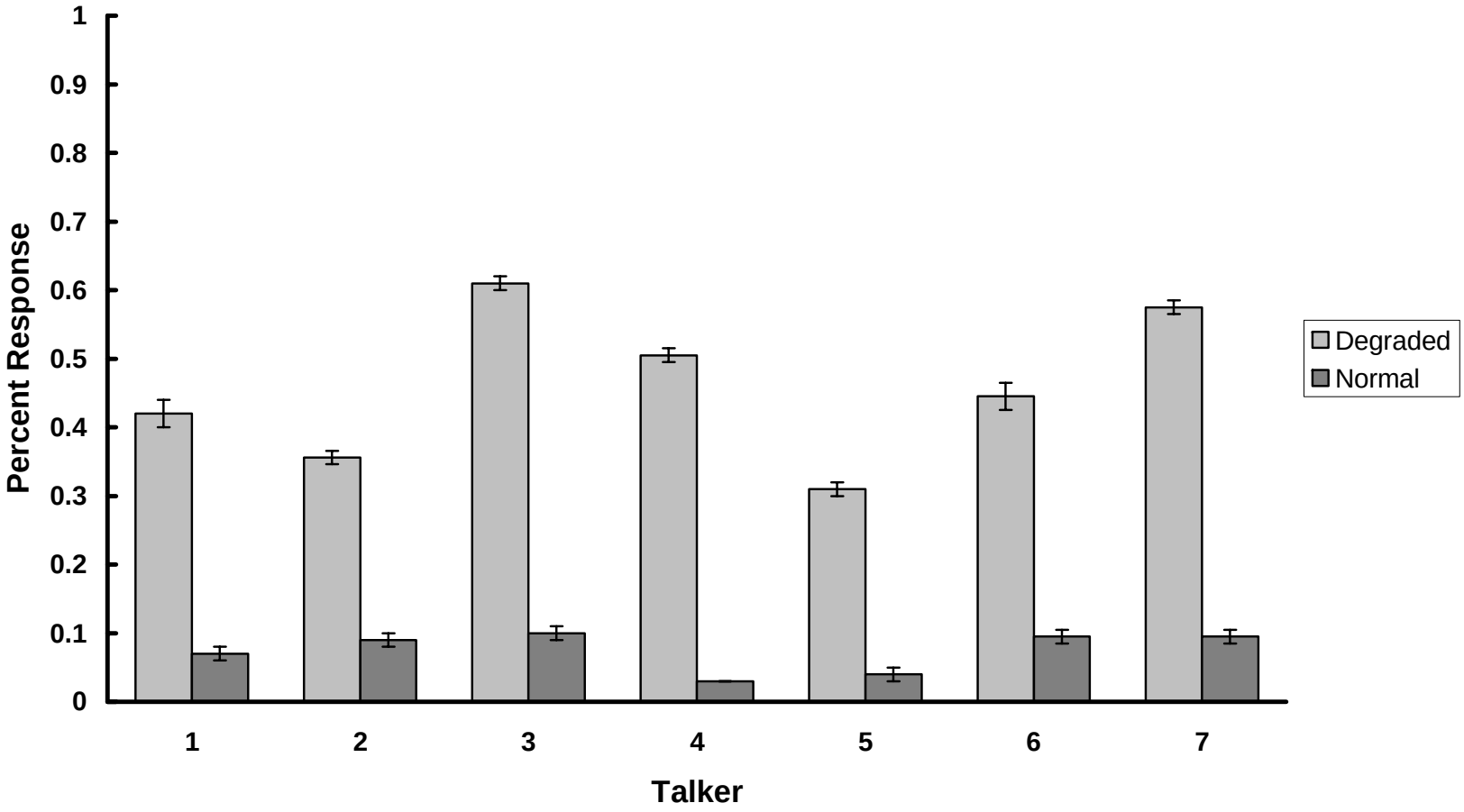
Percent Correct Identification Visual + Auditory --Degraded
Figure 3.



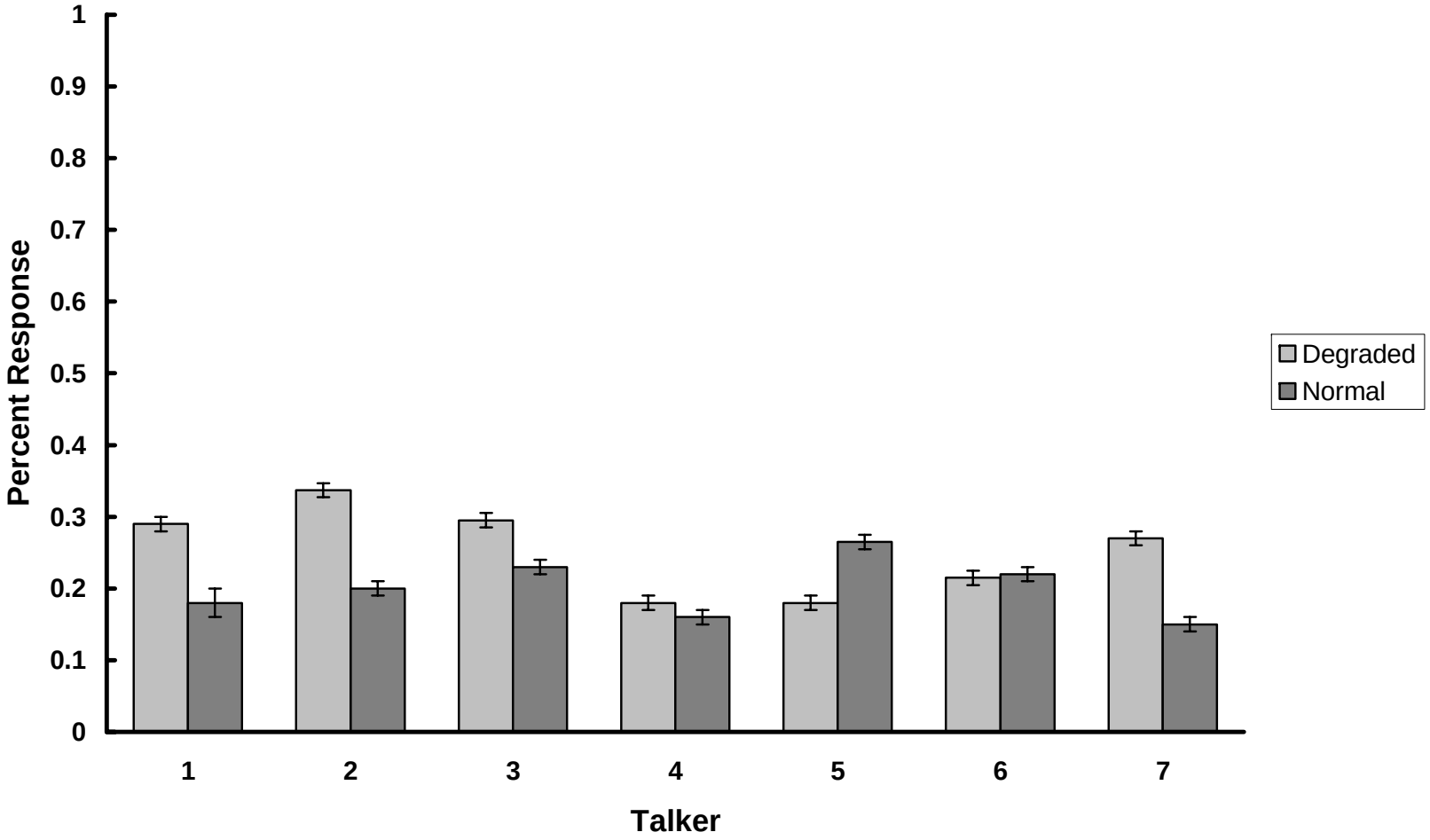
Percent "Auditory" Responses
Figure 4.



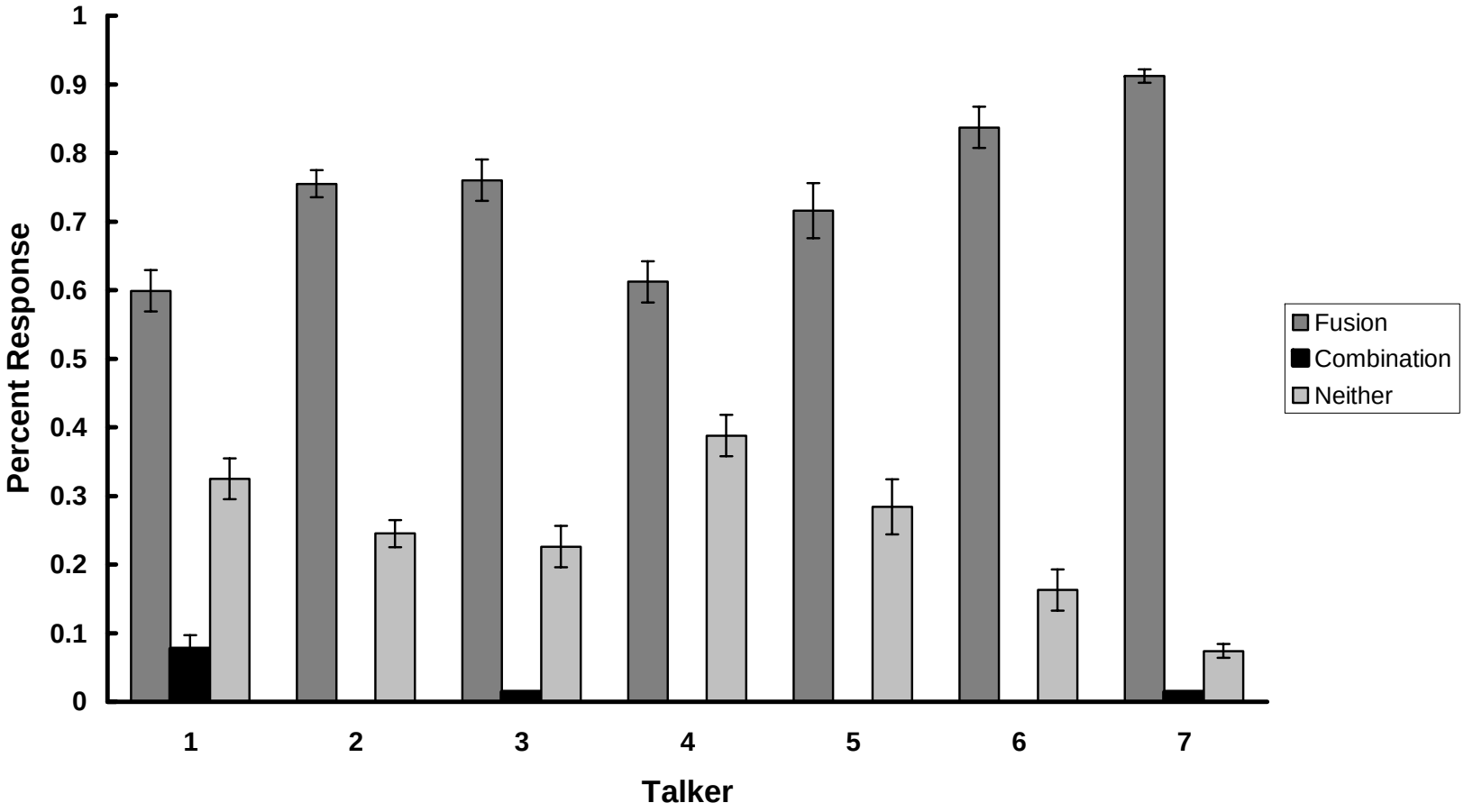
Percent "Visual" Responses
Figure 5.



Percent "Other" Responses
Figure 6.



Integration Response Types --Degraded
Figure 7.



Integration Response Types --Non-Degraded
Figure 8.